The effect of misclassification on estimates of relative risk

The abstract and article by Bennett et al in this issue of EBMH and the accompanying commentary by Jensen (p 93) make several important points. The article is a systematic review of studies that investigate whether early measures of externalising behaviour are able to predict later antisocial behaviour or conduct disorder. Bennett et al conclude that screening instruments have very poor positive predictive value; in other words, it is difficult to identify from kindergarten (children aged 4 and 5 years) reports of problem behaviour, those children likely to have antisocial behaviours several years later. In the accompanying commentary, Jensen points out that this has important implications for public policy and is somewhat at odds with the prevailing view on the stability of antisocial behaviour. The objective of this notebook article is to examine further one of the points of the review by Bennett et al as it has wider implications for the field of mental health in general.

There are at least 3 possible reasons for the low predictive ability of early screening instruments in this context. Firstly, there could be true lack of stability between externalising behaviours in kindergarten and conduct disorder several years later. Such early behaviours could be highly transient and not represent enduring traits of psychopathology. Secondly, it is possible that the screening instrument may not be able to pick up true cases of externalising behaviour in kindergarten and that there are a lot of “false positives” included in the high risk classification; that is, some of those without true externalising behaviour are being classified as high risk. Thirdly, it is possible that there is similar misclassification at the outcome point; that is, that true cases of conduct disorder are either not diagnosed or there are too many false positives.

Many readers might assume that the first explanation is most likely or that the other 2 possibilities would have little impact. It is easy to underestimate the dramatic effects that misclassification can have on longitudinal studies of this sort. I will illustrate below how a moderate degree of misclassification can have a dramatic effect on estimates of relative risk so that even given high levels of true stability, small relative risks will be observed (for a more general discussion, see reference 1).

There are 2 types of misclassification, differential and non-differential. In our example, non-differential misclassification occurs when the amount of measurement error in classifying a child as high or low risk is the same in those with and without the outcome variable. Thus, non-differential misclassification occurs when the degree of measurement error is the same in the group with and without later conduct disorder. It is random error, equally distributed among all observations. Differential misclassification occurs when the degree of measurement error is different in these groups; it is a systematic form of bias. In family history studies, for example, the ability of informants to correctly identify relatives as affected with some type of psychiatric disorder is greater if the informant is the relative of a “case” as opposed to the relative of a “control.” The effect of differential misclassification on estimates of relative risk is somewhat more complicated, and will be the topic of a future notebook.

Let me illustrate the effect of non-differential misclassification on hypothetical data that one might observe in the studies reviewed by Bennet et al. Let us assume that we have an inception cohort of 1000 children in kindergarten. We give a screening instrument to identify a high risk group; 200 children with high levels of externalising behaviour are identified, leaving us with 800 low risk children. Let us assume further that the true prevalence of conduct disorder several years later is 16.8%. The figure shows the $2 \times 2$ table illustrating a hypothetical association between predictor and outcome. The true relative risk in this example is 10 ($120/200 \div (48/800)$) and the true positive predictive value is 60% ($120/200$).

In reality, however, our measurement of externalising behaviour is not perfect. Some true high risk children will be missed because the screening instrument does not have perfect sensitivity. Some true low risk children will be misclassified as high risk because the instrument does not have perfect specificity. Let us assume that the sensitivity and specificity of the screening instrument is 80% or that the rate of false negatives and false positives is 20%. What effect will this have on the observed classification of individuals as high and low risk and on the subsequent estimates of relative risk and positive predictive value?

The table shows the impact of this relatively minimal error rate. The observed classification of children with and without later conduct disorder is presented as a function of their true risk classification and the sensitivity and specificity of the screening instrument. Among children with conduct disorder, our

<table>
<thead>
<tr>
<th>CD+</th>
<th>CD-</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR</td>
<td>120</td>
</tr>
<tr>
<td>LR</td>
<td>48</td>
</tr>
</tbody>
</table>

true prev = 16.8%  
true spec = 90%  
true PPV = 60%

Figure Risk status and later conduct disorder.
Effect of misclassification of risk status

<table>
<thead>
<tr>
<th>Classified</th>
<th>True</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CD+</td>
<td>HR</td>
<td>a</td>
</tr>
<tr>
<td>CD−</td>
<td>LR</td>
<td>b</td>
</tr>
</tbody>
</table>

Observed PPV = 106/320 = 33% (a + b) / a+b+g+h

Observed RR = 3.6 [ (a+b) / (a+b+e+f) ] / [c+d+c+e+d+g+h]

CD + : children with true conduct disorder
CD − : children without true conduct disorder
LR: low risk
HR: high risk

screening instrument will misclassify 20% of the true high risk children as low risk (0.2 × 120 = 24), and furthermore, 20% of the low risk children (0.2 × 480 = 96 or 10 children) will be misclassified as high risk. Similarly, among the children without conduct disorder 16 (0.2 × 80) and 150 (0.2 × 752) children will be misclassified. Thus, the observed classification of children will be the sum of the true and misclassified children: 320 high risk (106 + 214) and 680 (62 + 618) low risk children. This is in contrast with 200 and 800 children in the true situation.

What impact will this have on relative risk, in predicting antisocial behaviour at time 2? The rate of conduct disorder among the high risk group is now 33% (106/320) and the rate of conduct disorder among the low risk group is 9.1% (62/680). Thus, the observed relative risk in this example is now 3.6. This is quite a dramatic reduction in the relative risk from a value of 10 when the classification of risk status was perfect. The positive predictive value has also been reduced from the true value of 60% to an observed value of 33%. These values are not too dissimilar from those reported in the systematic review by Bennet et al.

The key issue is that in the presence of non-differential misclassification, the true relative risk will always be biased downwards. The magnitude of the bias depends on sensitivity, specificity, and the prevalence of the outcome (in our example, the prevalence of antisocial behaviour in the low risk group). The lower the prevalence, the greater the impact of the misclassification of the 800 low risk individuals.

The example we have provided has said nothing about the impact of misclassification at the level of outcome. There is no reason to believe that the studies reviewed by Bennet et al. used instruments to measure conduct disorder with perfect sensitivity and specificity. The effect of misclassification of antisocial behaviour is entirely analogous to the effect of misclassification of risk. This is because the calculation of relative risk is completely symmetric whether misclassification is at the level of risk factor or the outcome. Misclassification of both will, of course, have a cumulative effect. It is possible to show that, given the same sensitivity and specificity of 80% when measuring antisocial behaviour, the relative risk is observed to be less than 2.

Bennet et al. suggest that for screening to be effective, a more detailed measurement study must be undertaken of kindergarten children to improve sensitivity and specificity. Actually, such measurements do not need to be taken on all children but only on a subsample. It is possible to correct the observed relative risk back to its true value if the investigator knows the sensitivity and specificity of the screening instrument.

The effect of non-differential misclassification on estimates of relative risk will be to always reduce the true relative risk. The effect will be similar in studies of causation and treatment; in any study, in fact, where measurement is an issue and strength of association is important. Differential misclassification will influence the relative risk in either direction and to that extent is a more serious threat to interpretation. There is clearly a great need for studies in the mental health field to pay great attention to these issues and for readers of research articles to appreciate the ubiquitous effect that error has on our ability to apply the results of empirical data to patient care.

PETER SZATMARI, MD
Editor, Evidence-Based Mental Health

How to cite Evidence-Based Mental Health

CITATION OF MATERIAL FROM THE NOTEBOOK:

CITATION FOR MATERIAL TAKEN FROM A STRUCTURED ABSTRACT, WRITTEN WITHOUT ATTRIBUTION BY A STAFF MEMBER:

CITATION FOR MATERIAL TAKEN FROM A COMMENTARY TO AN ARTICLE: